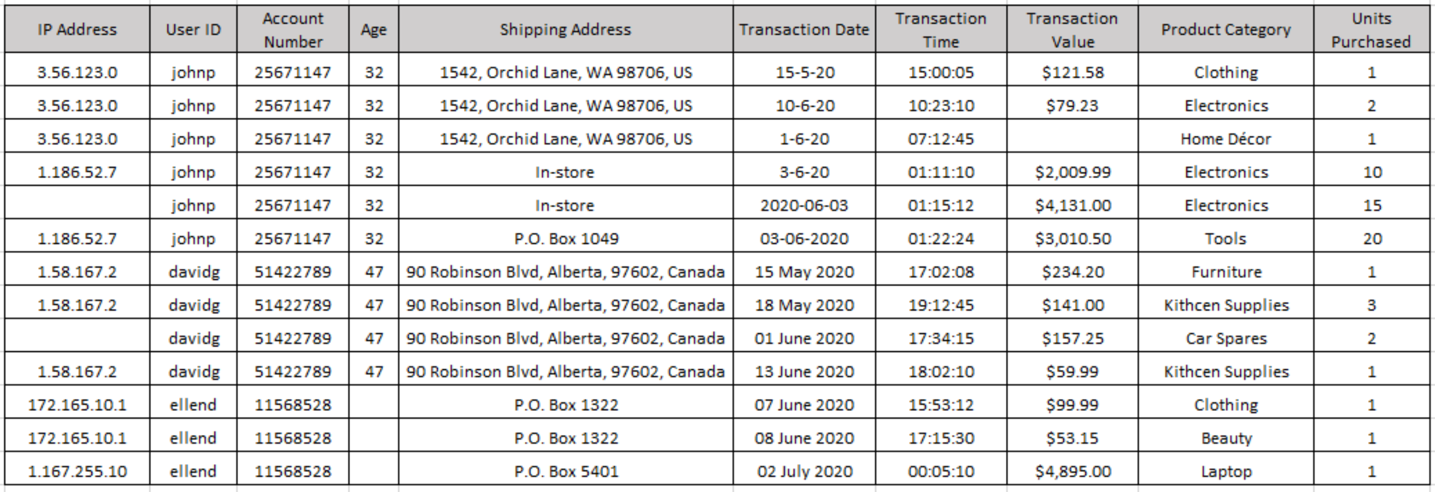
Project: Using Data Analysis for Detecting Credit Card Fraud

Prompt:

Given a dataset of three users' transactions, use data analysis techniques to detect possible fraudulent credit card activity.

Sample data set that captures the credit card transaction details for a few users:



IP Address	User ID	Account Number	Age	Shipping Address	Transaction Date	Transaction Time	Transaction Value	Product Category	Units Purchased
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	15-5-20	15:00:05	$121.58	Clothing	1
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	10-6-20	10:23:10	$79.23	Electronics	2
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	1-6-20	07:12:45		Home Décor	1
1.186.52.7	johnp	25671147	32	In-store	3-6-20	01:11:10	$2,009.99	Electronics	10
	johnp	25671147	32	In-store	2020-06-03	01:15:12	$4,131.00	Electronics	15
1.186.52.7	johnp	25671147	32	P.O. Box 1049	03-06-2020	01:22:24	$3,010.50	Tools	20
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	15 May 2020	17:02:08	$234.20	Furniture	1
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	18 May 2020	19:12:45	$141.00	Kithcen Supplies	3
	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	01 June 2020	17:34:15	$157.25	Car Spares	2
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	13 June 2020	18:02:10	$59.99	Kithcen Supplies	1
172.165.10.1	ellend	11568528		P.O. Box 1322	07 June 2020	15:53:12	$99.99	Clothing	1
172.165.10.1	ellend	11568528		P.O. Box 1322	08 June 2020	17:15:30	$53.15	Beauty	1
1.167.255.10	ellend	11568528		P.O. Box 5401	02 July 2020	00:05:10	$4,895.00	Laptop	1

Analysis Process:

1- Data points of relevance to use case:

- **Card Holder's Details**: Name, age, billing address.
- **Transaction Details**: Date, time, transaction value, description of the purchase.
- **Delivery Details**: Shipping address, delivery method (e.g., in-store pickup, home delivery).
- **Location Data**: Geographic location of the transaction, which can be inferred from the billing address, shipping address, or IP address.
- **Network Information**: IP address, device ID used for the transaction.
- **Transaction History**: Previous transactions to establish behavior patterns.

2- Data cleaning:

The data cleaning process involves several key actions to ensure the data is accurate and suitable for analysis:

- **Handling Missing Values**: Identify transactions with missing data points, such as transaction values. Decisions on handling these might include omitting the transaction from certain analyses (as with the omitted **john** transaction), imputing a value based on certain criteria, or flagging these transactions for further review.
- **Correcting Errors**: Identify and correct inaccuracies, such as typos in the card holder's details or transaction details, ensuring data integrity.
- **Standardizing Data Formats**: Ensure that all data points follow a consistent format, such as standardizing date fields to a uniform format (DD-MM-YYYY or YYYY-MM-DD), which is crucial for chronological analysis.
- **Identifying and Removing Duplicates**: Remove or consolidate duplicate records that could skew the analysis.
- **Outlier Detection and Handling**: Identify transactions that significantly deviate from the rest of the data. Decide on a case-by-case basis whether these outliers represent fraudulent transactions or legitimate but unusual activities.

Result(s) of data cleaning:

John's 3rd transaction was omitted from further analysis as it had no transaction value, which is essential for any sort of analysis. This value can't be assumed.

3- Data Visualization:

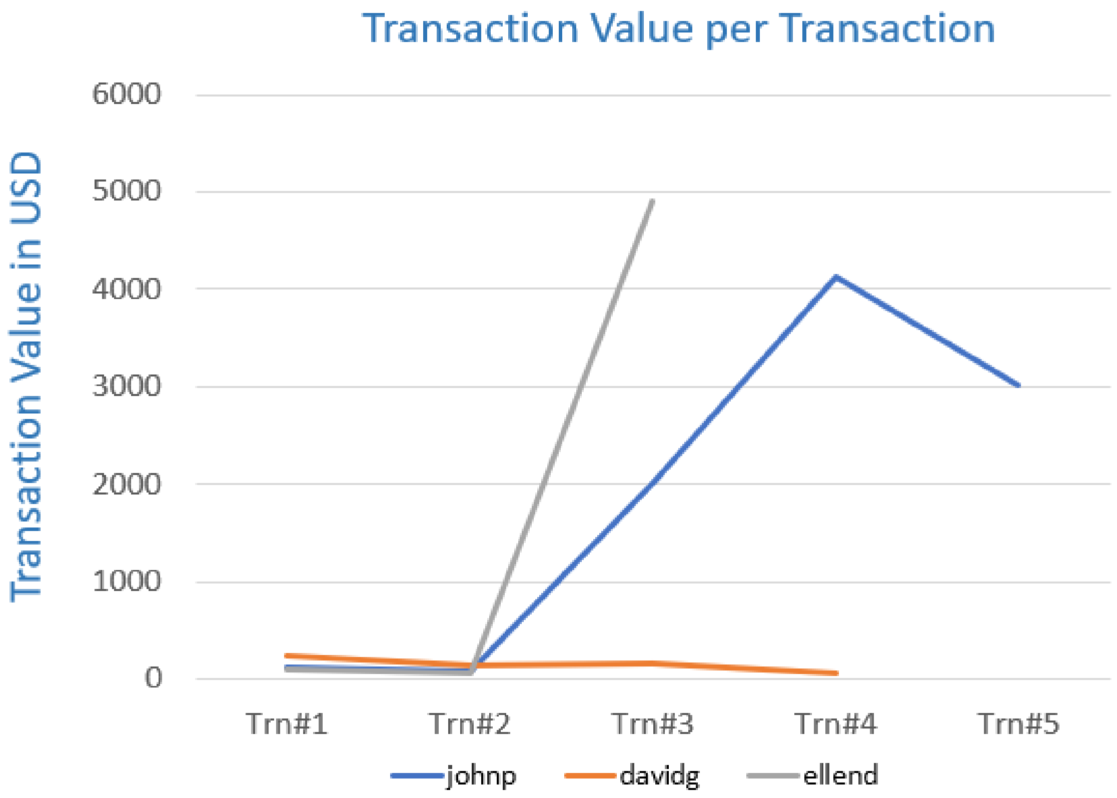
A line chart was chosen for this case study to illustrate changes and trends in transaction values over a sequence of transactions. Line charts are particularly effective in displaying data over time or in a sequential order, as they allow viewers to easily see the progression or movement between data points.

In the context of analyzing transaction values, a line chart helps to:

- **Track Trends**: It shows how transaction values change from one transaction to the next, making it easier to spot overall trends such as increasing or decreasing transaction amounts.

- **Identify Anomalies**: Spikes or dips in the line can quickly draw attention to outliers, such as transactions that are significantly higher or lower than the surrounding data points, which may indicate potential fraudulent activity.
- **Compare Multiple Entities**: The chart can display multiple lines, each representing a different individual's transaction values, allowing for a direct comparison across different data sets on the same scale.

Overall, the line chart provides a clear visual representation of the data that can be quickly interpreted to gain insights into spending patterns and to spot any irregularities that may warrant further investigation.



Questions:

1. **List at least 5 (five) data points that are required for the analysis and detection of a credit card fraud. (3 marks)**

    1. **Transaction Value**: High-value transactions or transactions that significantly deviate from a user's typical spending pattern can indicate potential fraud.

   2. **Transaction Frequency**: A sudden increase in the number of transactions over a short period, particularly if it's uncharacteristic of the user's normal behavior.

   3. **Shipping Address**: Inconsistencies or changes in shipping addresses, especially to locations that do not match the customer’s billing address or past delivery preferences.

   4. **IP Address**: An IP address that is inconsistent with the customer’s usual IP range or geolocation can suggest that a transaction may not be legitimate.

   5. **Transaction Date and Time**: Transactions that occur at odd hours, or on dates that don't align with the customer’s usual purchasing patterns, may warrant additional scrutiny.

2. **Identify 3 (three) errors/issues that could impact the accuracy of your findings, based on a data table provided. (3 marks)**

   1. **Incomplete Data**: If the dataset does not include all the transactions for each user, it would not accurately reflect their purchasing behavior, leading to potential false positives or negatives in fraud detection.

   2. **Missing Information**: Fields such as IP address, transaction value, or shipping address are missing or empty in some transactions. This lack of information can hinder the ability to fully analyze the transaction for potential fraud.

   3. **Data Entry Errors**: Incorrectly entered transaction details, such as a wrong shipping address, IP address, or transaction value, can lead to misidentification of a transaction as fraudulent or legitimate.

3. **Identify 2 (two) anomalies, or unexpected behaviors, that would lead you to believe the transaction may be suspect, based on a data table provided. (2 marks)**

   1. **The Frequency and Timing of johnp's Transactions**: Multiple transactions, including those high-value ones, are made in very close succession on the same day. This is a red flag as it indicates an unusual burst of activity that is atypical for the transaction history provided.

   2. **Ellend**'s transaction on 02 July 2020 is suspect for several reasons:

1. There is a dramatic increase in transaction value to $4,895.00, which is a significant spike from the previous 2 transactions.
2. This higher-value transaction was for a laptop, a departure from the previous purchase of clothing and beauty products.
3. The transaction was shipped to a different P.O. Box address (to the previous two transactions), which, while not inherently suspicious, can be a method used to obscure the recipient's actual location and is often scrutinized in fraud analysis.
4. This transaction is associated with a different IP address, **1167.255.10**, which differs from the two previous IP addresses. The change in IP address could suggest that the purchase was made from a different location, adding another layer of potential risk, especially if the IP geolocation significantly differs from **ellend**'s usual transactions.
5. It was the final transaction, which is a behavior typical for fraudsters who only use credit cards for one big transaction. Multiple and consequent big transactions could be assumed to be fraud and could be blocked by the bank.
6. Very late purchase at 00:05:10 which is also an unusual timing in comparison to ellend's previous 2 transactions.

4. Briefly explain your key take-away from the provided data visualization chart. (1 mark).

Neglecting additional information in the data set and by analyzing the line chart, what stands out is ellend's 3rd transaction showing a spike in value from $99.99 and $53.15 to $4.8950.00. Moreover, there are no additional transactions to show any sort of trend (outlier) and is a typical behavior for fraudsters.

5. Identify the type of analysis that you are performing when you are analyzing historical credit card data to understand what a fraudulent transaction looks like. [Hint: The four types of Analytics include: Descriptive, Diagnostic, Predictive, Prescriptive] (1 mark0.

- For the creation of a sample dataset and a line chart, **descriptive analysis** was used. Descriptive analytics is crucial for setting the stage, providing a clear picture of the transaction landscape before delving into deeper analysis.
- However, when using both the dataset table and the line chart’s data to examine the data to understand why certain transactions stand out as potential indicators of fraud, **diagnostic analysis** was used.